IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), New York City,June 16-22, 2006

Toward Robust Distance Metric Analysis for Similarity Estimation

Jie Yu[1], Jaume Amores[2], Nicu Sebe[3], Qi Tian[1]
[1]Department of Computer Science, University of Texas at San Antonio, TX, USA
{jyu, qitian}@cs.utsa.edu
[2]IMEDIA Research Group, INRIA, Rocquencourt, France
jaume.amores@inria.fr
[3]Faculty of Science, University of Amsterdam, Netherlands
nicu@science.nva.nl

Abstract

In this paper, we present a general guideline to establish the relation between a distribution model and its corresponding similarity estimation. A rich set of distance metrics, such as harmonic distance and geometric distance, is derived according to Maximum Likelihood theory. These metrics can provide a more accurate feature model than the conventional Euclidean distance (SSD) and Manhattan distance (SAD). Because the feature elements are from heterogeneous sources and may have different influence on similarity estimation, the assumption of single isotropic distribution model is often inappropriate. We propose a novel boosted distance metric that not only finds the best distance metric that fits the distribution of the underlying elements but also selects the most important feature elements with respect to similarity. We experiment with different distance metrics for similarity estimation and compute the accuracy of different methods in two applications: stereo matching and motion tracking in video sequences. The boosted distance metric is tested on fifteen benchmark data sets from the UCI repository and two image retrieval applications. In all the experiments, robust results are obtained based on the proposed methods.

1. Introduction

In many science and engineering fields, the similarity between two features is determined by computing the distance between them using a certain distance metric. In computer vision as well as some other areas, the Euclidean distance or SSD (L_2 - sum of the squared differences) is one of the most widely used metrics. However, it has been suggested that it is not appropriate for many problems [1]. From a maximum likelihood perspective, it is well known that the SSD is justified when the feature data distribution is Gaussian [2] while the Manhattan distance or SAD (L_1 - sum of the absolute differences), another commonly used metric, is justified when the feature data distribution is Exponential (double or two-sided exponential). Therefore, one can determine which metric to use if the underlying data distribution is known or well estimated. The common assumption is that the real distribution should fit either the Gaussian or the Exponential. However, in many applications this assumption is invalid. Finding a suitable distance metric becomes a challenging problem when the underlying distribution is unknown and could be neither Gaussian nor Exponential.

In content-based image retrieval feature elements are extracted for different statistical properties associated with entire digital images, or perhaps with specific region of interest. The heterogeneous sources suggest that the elements may be from different distributions. In previous work, most of the attention focused on extracting low-level feature elements such as color [3], texture [4], and shape [5] with little or no consideration of their distributions. The most commonly used method for calculating the similarity between two feature vectors is still to compare the Euclidean distance between them.

Although some works have been done to utilize the data model in similarity image retrieval [2,6], the relation between the distribution model and the distance metric has not been fully studied yet. It has been justified that Gaussian, Exponential, and Cauchy distribution result in L_2 , L_1 and Cauchy metrics, respectively. However, distance metrics that fit other distribution models have not been studied yet. The similarity estimation based on feature elements from unknown distributions is an even more difficult problem. In this paper based on our previous work in [2, 6] we propose a guideline to learn a robust distance metric for accurate similarity estimation.

The rest of the paper is organized as follows. Section 2 presents a distance metric analysis using the maximum likelihood approach. Section 3 describes the boosted distance metric. In Sections 4 and 5 we apply the new distance metrics to estimate the similarity in a stereo matching application, motion tracking in a video sequence, and content-based image retrieval. Discussions and conclusions are given in Section 6.

2. Distance Metric Analysis

2.1. Maximum Likelihood Approach

The additive model is a widely used model in computer vision regarding maximum likelihood estimation. Haralick and Shapiro [7] consider this model in defining the M-estimate: "Any estimate μ defined by a minimization problem of the form $\min \sum_i f(x_i - \mu)$ is called an M-estimate." Note that the operation "–" between the estimate and the real data implies an additive model. The variable μ is either the estimated mean of a distribution or, for simplicity, one of the samples from that distribution.

Maximum likelihood theory [7] allows us to relate a data distribution to a distance metric. From the mathematical-statistical point of view, the problem of finding

the right measure for the distance comes down to the maximization of the similarity probability.

We use image retrieval as an example for illustration. Consider first, two subsets of N images from the database (D): $X \subset D$, $Y \subset D$ which according to the ground truth are similar:

$$X \equiv Y \text{ or } x_i \equiv y_i,\ i = 1,..,N \qquad (1)$$

where $x_i \in X$, $y_i \in Y$ represent the images from the corresponding subsets.

The equation (1) can be rewritten as:

$$x_i = y_i + d_i,\ i = 1,..,N \qquad (2)$$

where d_i represents the "distance" image obtained as the difference between image x_i and y_i .

In this context the similarity probability between two sets of images X and Y can be defined:

$$P(X,Y) = \prod_{i=1}^{N} p(x_i, y_i) \qquad (3)$$

where $p(x,y)$ is the probability density function between images x and y. Independence across images is assumed. We define

$$f(x_i, y_i) = -\log p(x_i, y_i) \qquad (4)$$

Then equation (3) becomes

$$P(X,Y) = \prod_{i=1}^{N} \{\exp[-f(x_i, y_i)]\} \qquad (5)$$

where the function f is the negative logarithm of the probability density function of images x and y.

According to (5) we have to find the function f that maximizes the similarity probability. This is the *maximum likelihood* estimator for X, given Y [7].

In the above considerations, we are talking about images but this notion can be extended to feature vectors associated with the images when we are working with image features or, even, can be extended to pixel values in the images. Taking the logarithm of (5) we find that we have to minimize the expression:

$$\sum_{i=1}^{N} f(x_i, y_i) \qquad (6)$$

In this case, according to equation (2) the function f does not depend individually on its two arguments, query image x_i and the predicated one y_i, but only on their difference. We have thus a local estimator and we can use $f(d_i)$ instead of $f(x_i, y_i)$ where $d_i = x_i - y_i$ and the operation "–" denotes pixel-by-pixel difference between the images, or an equivalent operation in feature space.

Therefore, minimizing equation (6) is equivalent to minimizing

$$\sum_{i=1}^{N} f(d_i) \qquad (7)$$

From the above discussion we find that *Maximum likelihood* estimation shows a direct relation between the data distribution and the comparison metric.

2.2. Distance Metric Analysis

The Gaussian, Exponential, and Cauchy distribution models result in the L_2 metric, L_1 metric, and Cauchy metric, respectively [2]. It is reasonable to assume that there may be other distance metrics that fit the unknown real distribution better. More accurate similarity estimation is expected if the metric could reflect the real distribution. We call this problem of finding the best distance metric *distance metric analysis*. It can be mathematically formulated as follows.

Suppose we have observations

$$x_i = \mu + d_i \qquad (8)$$

where d_i , $i = 1, \cdots, N$ are data components and μ is the distribution mean or a sample from the same class if it is considered as center of a subclass from a locality point of view. In most cases μ is unknown and may be approximated for similarity estimation. For some function

$$f(x, \mu) \geq 0 \qquad (9)$$

which satisfies the condition $f(\mu, \mu) = 0$, μ can be estimated by $\hat{\mu}$ which minimizes

$$\varepsilon = \sum_{i=1}^{N} f(x, \hat{\mu}) \qquad (10)$$

It is equivalent to satisfy

$$\sum_{i=1}^{N} \frac{d}{d\hat{\mu}} f(x_i, \hat{\mu}) = 0 \qquad (11)$$

For some specific distributions, the estimated mean $\hat{\mu} = g(x_1, x_2, \cdots, x_N)$ has a closed form solution. The arithmetic mean, median, harmonic mean, and geometric mean in Table 1 are in that category. It's well-known that the L_2 metric (SSD) corresponds to the arithmetic mean while the L_1 metric (SAD) corresponds to the median. However, no literature has discussed the distance metrics associated with the distribution models that imply the harmonic mean or the geometric mean. Those metrics in Table 1 are inferred using equation (11).

Table 1. Distance metrics and mean estimation for different distributions

	Distance Metric	Mean Estimation
Arithmetic	$\varepsilon = \sum_{i=1}^{N} (x_i - \hat{\mu})^2$	$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$
Median	$\varepsilon = \sum_{i=1}^{N} \lvert x_i - \hat{\mu} \rvert$	$\hat{\mu} = med(x_1, \cdots, x_N)$
Harmonic	$\varepsilon = \sum_{i=1}^{N} x_i (\frac{\hat{\mu}}{x_i} - 1)^2$	$\hat{\mu} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$
Geometric	$\varepsilon = \sum_{i=1}^{N} [\log(\frac{x_i}{\hat{\mu}})]^2$	$\hat{\mu} = (\prod_{i=1}^{N} x_i)^{\frac{1}{N}}$

Figure 1(a) illustrates the difference among the distance functions $f(x, \hat{\mu})$ for the arithmetic mean, median, harmonic mean and geometric mean. For fair comparison the value of μ is set to be 10 for all distributions. We found that in distribution associated with the harmonic and geometric estimations, the observations which are far away from the correct estimate (μ) will make less contribution in producing μ, as distinct from the arithmetic mean. In that case the estimated values will be less

sensitive to the bad observations (i.e., observation with large variance), and they are therefore more robust.

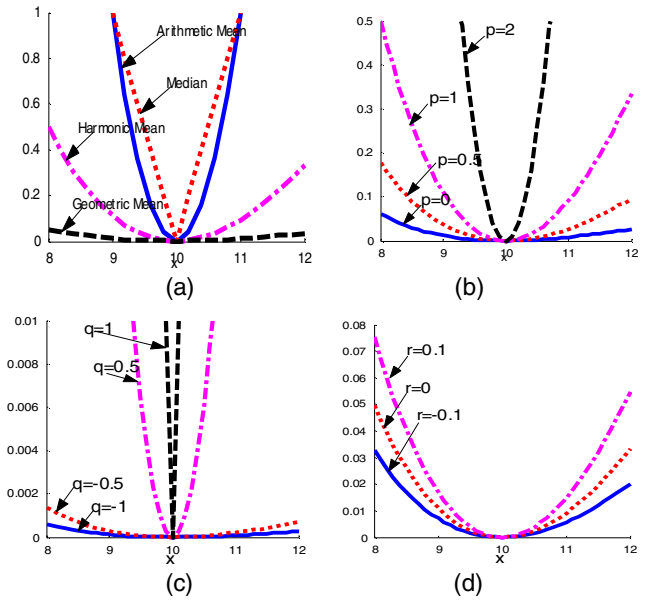


Figure 1. The distance function $f(x,\mu)$ of (a) the arithmetic mean, median, harmonic mean and geometric mean, (b) 1st-type, (c) 2nd-type generalized harmonic mean, and (d) the generalized geometric mean (μ is fixed and set to 10)

2.3. Generalized distance metric analysis

The robust property of Harmonic and Geometric distance metrics motivates us to generalize them and come up with new metrics that may fit the distribution better. Three families of distance metrics in Table 2 are derived from the generalized mean estimation using equation (10). The parameters p, q, r define the specific distance metrics and describe the corresponding distribution models which may not be explicitly formulated as Gaussian and Exponential.

We found that in the generalized harmonic mean estimation the 1st type is generalized based on the distance metric representation, while the 2nd type is generalized based on the estimation representation. However, if $p=1$ and $q=-1$, both types will become ordinary harmonic mean, and if $p=2$ and $q=1$, both types will become arithmetic mean. As for the generalized geometric mean estimation, if $r=0$, it will become an ordinary geometric mean. It is obvious that the generalized metrics correspond to a wide range of mean estimations and distribution models. Figures 1 (b)-(d) show the distance metric function $f(x,\hat{\mu})$ corresponding to the 1st type and 2nd type generalized harmonic mean, and the generalized geometric mean estimation, respectively. It should be noted that not all mean estimations have closed-form solution as in Tables 1 and 2. In that case $\hat{\mu}$ can be estimated by numerical analysis, e.g., greedy search of $\hat{\mu}$ to minimize ε.

Table 2. Generalized distance metrics

	Distance metric	Mean Estimation
Generalized harmonic (1st type)	$\varepsilon=\sum_{i=1}^{N}(x_i)^p(\frac{\hat{\mu}}{x_i}-1)^2$	$\hat{\mu}=\frac{\sum_{i=1}^{N}(x_i)^{p-1}}{\sum_{i=1}^{N}(x_i)^{p-2}}$
Generalized harmonic (2nd type)	$\varepsilon=\sum_{i=1}^{N}[(x_i)^q-(\hat{\mu})^q]^2$	$\hat{\mu}=[\frac{N}{\sum_{i=1}^{N}(x_i)^q}]^{-\frac{1}{q}}$
Generalized geometric	$\varepsilon=\sum_{i=1}^{N}[(x_i)^r\log(\frac{x_i}{\hat{\mu}})]^2$	$\hat{\mu}=[\prod_{i=1}^{N}(x_i)^{(x_i)^{2r}}]^{\frac{1}{\sum_{i=1}^{N}(x_i)^{2r}}}$

3. Boosting Distance Metrics for Similarity Estimation

3.1. Motivation

As we mentioned in Section 1, the most commonly used distance metric is the Euclidean distance that assumes the data have a Gaussian isotropic distribution. When the feature space has a large number of dimensions, an isotropic assumption is often inappropriate. Besides, the feature elements are often extracted by different statistical approaches, and their distributions may not be the same and different distance metric may better reflect the distribution. Thus, an anisotropic and heterogeneous distance metric may be more suitable for estimating the similarity between features.

Mahalanobis distance $(x_i-y_i)^TW(x_i-y_i)$ is one of the traditional anisotropic distances. It tries to find optimal estimation of the weight matrix W. It is worth noticing that it assumes the underlying distribution is Gaussian, which is often not true. Furthermore, if d is the number of dimensions, the matrix W contains d^2 parameters to be estimated, which may not be robust when the training set is small compared to the number of dimensions. One may suggest that we can apply classical techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) to reduce the dimensions. However, these methods cannot solve the problems of a small training set and they also assume Gaussian distribution.

3.2. Boosted Distance Metrics

Based on the analysis in Section 3.1, we propose a boosted distance metrics for similarity estimation where similarity function for certain class of samples can be estimated by a generalization of different distance metrics on selected feature elements. In particular, we use AdaBoost with decision stumps [8] and our distance metric analysis to estimate the similarity.

Given a training set with feature vectors x_i, the similarity estimation is done by training AdaBoost with differences between vectors $d=x_i-x_j$, where each difference vector d has an associated label l_d

$$l_d=\begin{cases}1 \text{ if } x_i \text{ and } x_j \text{ are from same class}\\ 0 \text{ otherwise}\end{cases}$$

A weak classifier is defined by a distance metric m on a

feature element f with estimated parameter(s) θ, which could be as simple as the mean and/or a threshold. The label prediction of the weak classifier on feature difference d is $h_{m,f,\theta}(d) \in \{0,1\}$.

The boosted distance metric $H(d)$ is learnt by weighted training with different distance metrics on each feature elements and selecting the most important feature elements for similarity estimation iteratively. Consequently we derive a predicted similarity $S(x,y) = H(x-y)$ that is optimal in a classification context. The brief algorithm is listed below.

Algorithm **Boosting Distance Metric**

Given:

A pair wise difference vector set D and the corresponding label L

Number of Iteration T

Weak classifiers based on each distance metric m for each feature element f

Initialization: weight $w_{i,t=1} = 1/|D|$

Boosting:

For $t=1,..,T$

- Train the weak classifier on the weighted sample set
- Select the best weak classifier giving the smallest error rate
  $$\varepsilon_t = \min_{m,f,\theta} \sum w_{i,t} \, | h_{m,f,\theta}(d_i) - l_i |$$
- Let $h_t = h_{m_t,f_t,\theta_t}$ with m_t, f_t, θ_t minimizing error rate
- Compute the weights of classifiers (α_t) based on its classification error rate
  Let $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$, $\alpha_t = \frac{1}{\log(\beta_t)}$
- Update and normalize the weight for each sample
  $$w_{i,t+1} = w_{i,t} \beta_t^{1-|h_{t,i}-l_i|}$$
  $$w_{i,t+1} = w_{i,t+1} / \sum_i w_{i,t+1}$$

end for t

Final prediction $H(d) = \sum_t \alpha_t h_t(d)$

The proposed method has three main advantages: i) the similarity estimation only uses a small set of elements that is most useful for similarity estimation; ii) for each element the distance metric that best fits its distribution is learnt; iii) it adds effectiveness and robustness to the classifier when we have a small training set compared to the number of dimensions. Since the training iteration T is usually much less than the original data dimension, the boosted distance metric works as a non-linear dimension reduction technique, which keeps the most important elements to similarity judgment. It could be very helpful to overcome the small sample set problem. It is worth mentioning that the proposed method is general and can be plugged into many similarity estimation techniques, such as the widely used K-NN.

3.3. Related work

We notice that there have been several works on estimating the distance to solve certain pattern recognition problems. Domeniconi *et al.* [9] and Peng *et al.* [10] propose specific estimations designed for the K-NN classifier. They obtain an anisotropic distance based on local neighborhoods that are narrower along relevant dimensions and more elongated along non-relevant ones. Xing *et al.* [11] propose to estimate the matrix W of a Mahalanobis distance by solving a convex optimization problem. They apply the resulting distance to improve the K-means behavior. Bar-Hillel et al. [12] also use a weight matrix W in order to estimate the distance by Relevant Component Analysis (RCA). They improve the Gaussian Mixture EM algorithm by applying the estimated distance along with equivalence constraints. The work by Hertz *et al.* [13] resembles the boosting part of our method, although it is conceptually different. They use AdaBoost to estimate a distance function in a product space (with pairs of vectors) while the weak classifier minimizes an error in the original feature space. Therefore, the weak classifier minimizes a different error than the one minimized by the strong classifier AdaBoost. In contrast, our framework utilizes AdaBoost with weak classifiers that minimize the same error as AdaBoost, and in the same space.

Compared to the previous work our proposed distance estimation method can be considered novel due to the following reasons: i) it is general for any kind of classifier (i.e. not specific for K-NN); ii) it learns the distance metric that fits the underlying distribution, instead of assuming it as Gaussian as in the Mahalanobis family metric; iii) it is not sensitive to small sample set problem. In summary, our method provides an effective learner that estimates an optimal distance given labeled data, and at the same time it does not assume any particular distribution of the data.

4. Experiments and Analysis

4.1. Distance Metric Analysis in Stereo Matching

Stereo matching is a computer vision application that aims at finding correspondences between entities in images with overlapping scene content. The images are typically taken from cameras at different viewpoints, so the intensities of corresponding pixels are often different. In this case automatic stereo matcher is expected to detect the corresponding point pairs registered in stereo images of the test set scenes. To compare different distance metrics we implement a stereo matcher based on the distance between different image regions. The optimal metric in this case will give the most accurate stereo matching performance.

Two standard stereo data sets, Castle and Tower, from Carnegie Mellon University are used for training and testing. In each image there are specific points labeled with ground truth information on correspondence relations. We want to find the position of ground truth points of one image in its consecutive images based on similarity estimation. An example of two stereo images from the Castle data set is given in Figure 2.

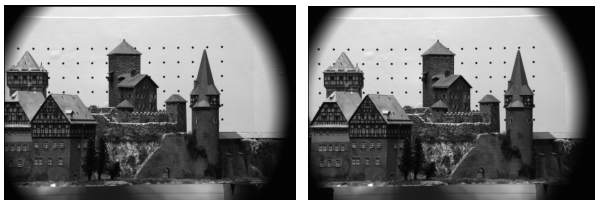


Figure 2. A stereo image pair from the Castle data set

Our stereo matcher tries to match a template defined around one point from an image with the templates around points in the other images in order to find similarity. If the resulting points are equivalent to those provided by the ground truth we consider that we have a *hit*, otherwise, we have a *miss*. The hit rate is given by the number of the hits divided by the number of possible hits (number of corresponding point pairs).

Using the distance metrics analysis discussed in Section 2, we obtain distances between pixel values of two templates. A set of stereo matchers can be constructed by finding a matching template that gives the smallest sum of distance based on different metrics. For the parameterized metric we search for the optimal value. The parameters of p, q, and r are tested in the range of -5 to 5 with step size 0.1.

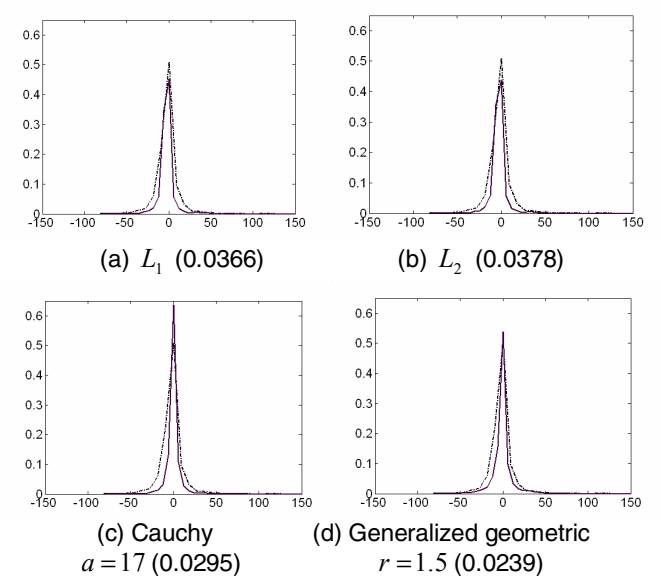


Figure 3. The real data distribution (dashed line) vs. the estimated data distribution (solid line) for different metric on Castle data set

Although empirical result may show one distance metric offer best performance, we want to verify if the ground truth distance matches the modeled distribution most accurately using Chi-Square test [14]. Figure 3 shows the real data distribution and the estimated data distribution for the distance metrics on the Castle data set. The largest hit rate and the smallest Chi-square test value are highlighted in bold. Both the solid and dashed curves are sampled with 233 points at the equal intervals. The Chi-square test values for each metric are shown in Table 3. The smaller the Chi-square test value, the closer the estimation is to the real distribution. The generalized geometric mean metric has the best fit to the measured data distribution. Therefore, one expects the accuracy to be the greatest when using the generalized geometric mean metric (Table 3). In all cases, the hit rate for the generalized geometric mean ($r = 1.5$) is 80.4%, and the hit rate for Cauchy metric is 78.9%. The hit rates obtained with L_1 and L_2 are both 78.2%. The Cauchy metric performs better than both L_1 and L_2. It should be noted here that the Chi-square test score is not exactly in the same order of the hit rate though the winner is consistent in both cases. This is because the ground truth is provided with subpixel accuracy during the data collection process, and we consider it is a hit when the corresponding point lies in the neighborhood of one pixel around the point provided by the ground truth. The inconsistency introduced by this rounding distance may explain the observation (not in the exact order for both measures). Similar results are obtained for the Tower set and are not shown here due to the limited available space.

Table 3. The accuracy (percent) of the stereo matcher on the Castle set (best parameter is shown)

Distance Metric	Chi-square test	Hit rate (%)
L_1	0.0366	78.2
L_2	0.0378	78.2
Cauchy	0.0295 ($a = 17$)	78.9 ($a = 17$)
Harmonic mean	0.0273	78.2
Geometric mean	0.0378	77.1
1st-type generalized harmonic mean(1st-gh)	0.0328 ($p = 1.5$)	78.2 ($p = 1.5$)
2nd-type generalized harmonic mean(2nd-gh)	0.0272 ($q = 1.6$)	78.6 ($q = 1.6$)
Generalized geometric mean(gg)	**0.0239** ($r = 1.5$)	**80.4** ($r = 1.5$)

4.2. Distance Metric Analysis in Motion Tracking

In this experiment distance metric analysis is tested on a motion tracking application. We use a video sequence containing 19 images of a moving head in a static background [15]. For each image in this video sequence, there are 14 points given as ground truth. The motion tracking algorithm between the test frame and another frame performs template matching to find the best match in a 5×5 template around a central pixel. In searching for the corresponding pixel, we examine a region of width and the height of 7 pixels centered at the position of the pixel in the test frame. The idea of this experiment is to trace moving facial expressions. Therefore, the ground truth points are provided around the lips and the eyes, which are moving through the sequences.

In figure 4, we display the fit between the real data distribution and the four distance metrics. The real data distribution is calculated using the template around points in the ground truth data set considering sequential frames. The best fit is the generalized geometric mean metric with $r = 7.0$.

Between the first frame and a later frame, the tracking distance represents the average template matching results. Figure 5 shows the average tracking distance of the different

distance metrics. The generalized geometric mean metric with $r = 7.0$ performs best, while Cauchy metric outperforms both L_1 and L_2.

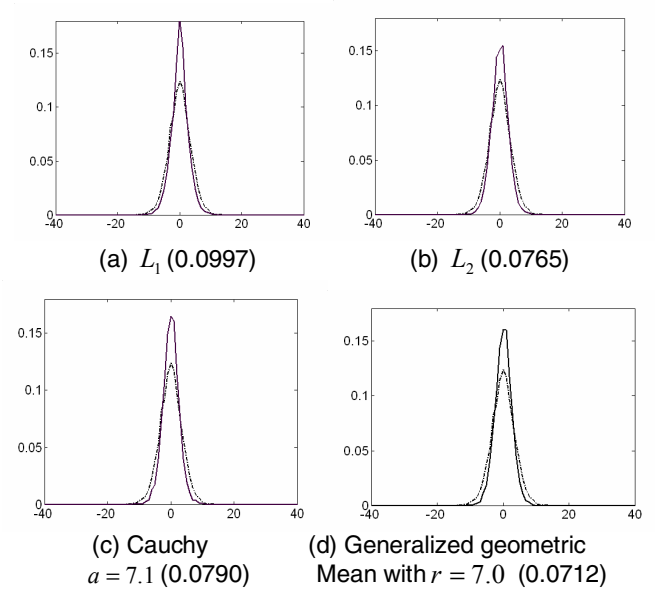


Figure 4. The real data distribution (dashed line) vs. the estimated data distribution (solid line) for motion tracking

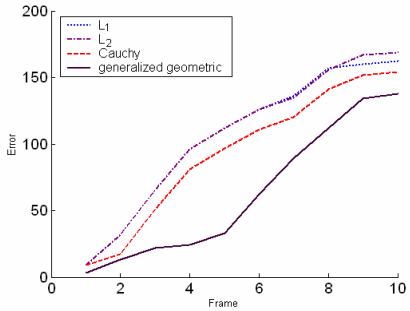


Figure 5. Average tracking distance of the corresponding points in successive frames

4.3. Boosted Distance Metric on Benchmark Data Set

In this section we compare the performance of our boosted distance metric with several well-known traditional approaches. The experiment is conducted on 13 benchmark datasets from UCI and two image retrieval datasets: a subset of MNIST data set [16], containing similar hand-written 1's and 7's and a gender recognition database containing facial images from the AR database [17] and the XM2TVS database [18]. The traditional distance metrics we tested are: L_1, L_2, RCA distance (RCA-D), Mahalanobis distance (Mah-D) and their variants (RCA-CD, Mah-CD). To make the comparison complete, we also test original AdaBoost with decision stump *(d.s.)* and C4.5. Due to the space limitation, only the traditional distance metric that gives the best performance in each data set is shown. The smallest error rates are highlighted in bold.

From the results in Table 4 we can find that our boosted distance metric performs the best in 12 out of 15 datasets. It provides comparable results to the best performance on 2 datasets. Only in one dataset our method is outperformed by the traditional distance metric. It proves that our method could discover the best distance metric that reflects the distribution and selects the feature elements that are discriminant in similarity estimation.

Table 4. Comparison to traditional distance metric and AdaBoost on UCI datasets

Error Rate (%)	Traditional Metric	AdaBoost +d.s.	AdaBoost +C4.5	Boosted Metric
ad	17.31 (L_1)	12	11.42	**8.88**
gender	15.38 (L_1)	12.27	11.89	**10.45**
mnist	2.34 (RCA-D)	2.22	2.14	**1.6**
arrhythmia	37.02 (RCA-D)	31.39	29.94	**25.78**
splice	10.55 (Mah-D)	5.94	4.84	**4.73**
sonar	26.1 (Mah-CD)	25.95	25.81	**25.67**
spectf	31.16 (Mah-D)	28.65	27.18	**27.1**
ionosphere	**10.78** (RCA)	19.92	19.92	16.27
wdbc	6.83 (Mah-CD)	5.81	5.37	**4.67**
german	38.74 (Mah-D)	34.31	33.18	**32.4**
vote1	9.07 (L_1)	**6.37**	6.37	6.86
credit	19.18 (Mah-CD)	17.97	**17.21**	18.33
wbc	5.25 (RCA)	5.7	5.34	**3.79**
pima	34.55 (Mah-CD)	31.02	29.96	**28.91**
liver	41.11 (Mah)	35.51	35.43	**33.58**

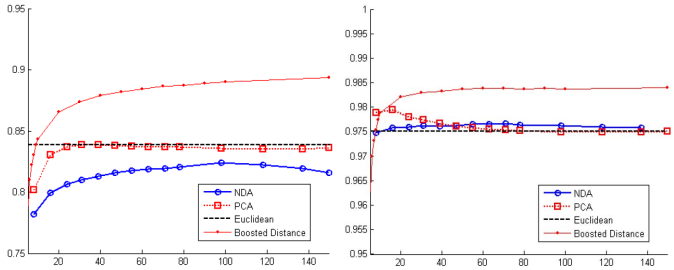
4.4. Boosted Distance Metric in Image Retrieval

As we discussed in Section 3.2, the boosted distance metric performs an element selection that is highly discriminant for similarity estimation and it doesn't suffer from the small sample set problem as LDA and other dimension reduction techniques. To evaluate the performance, we tested the boosted distance metric on image classification against some state-of-the-art dimension reduction techniques: PCA, LDA, NDA and plain Euclidean distance in the original feature space.

The two data sets we used are two image retrieval data sets same as in Section 4.3. The dimension of the feature for both databases is 784 while the size of training set is fixed to 200 which is small compared to the dimensionality of feature. In such circumstance, appropriate distance metric is very important. To play fair, simple Nearest-Neighbor classifier is used in the reduced dimension space.

Figure 6 shows the classification accuracy against the projected dimension, which, for our boosted distance metric, is the training iteration T. Because of the small sample problem, LDA has poor accuracy of 50% and 49.9% and is not shown in the figure. Simple regularization scheme may improve its

performance but still remains much worse than other techniques. It is clear that the traditional methods perform poorly due to the fact that we use a very small training set compared to the dimensionality of the data and all traditional methods rely on estimating a covariance or scatter matrix. Our boosted distance metric only needs to estimate very few parameters on each dimension, which provides a robust performance on the small training set and make it outperform the well-known techniques.



(a) (b)

Figure 6. Accuracy of classification on gender recognition (a) and hand written digits (b)

5. Discussions and Conclusions

This paper extends our work in [2,6] by comprehensive analysis on distance metric and boosting heterogeneous metrics for similarity estimation. Our main contribution is to provide a general guideline for designing robust distance estimation that could adapt data distributions automatically. Novel distance metrics deriving from harmonic, geometric mean and their generalized forms are presented and discussed. We examined the new metrics on several applications in computer vision, and the estimation of similarity can be significantly improved by the proposed distance metric analysis.

The relationships between probabilistic data models, distance metrics, and ML estimators have been widely studied. The creative component of our work is to start from an estimator and perform reverse engineering to obtain a metric. In this context, the fact that some of the proposed metrics cannot be translated into a known probabilistic model is both a curse and a blessing. A curse, because it is really not clear what the underlying probabilistic models are (they certainly do not come from any canonical family), and this is usually the point where one starts from. After all, the connection between the three quantities (metric, data model, ML estimator) is probabilistic. It is a bit unsettling to have no idea of what these models are. A blessing, is because this is probably the reason why these metrics have not been previously proposed. But they seem to work very well according to the experimental result in this paper.

In similarity estimation the feature elements are often from heterogeneous sources. The assumption that the feature has a unified isotropic distribution is invalid. Unlike traditional anisotropic distance metric, our proposed method does not make any assumption on the feature distribution. Instead it learns distance metrics on each element to capture the underlying feature structure. Because the distance metrics are trained on the observations of each element, the boosted distance does not suffer from the small sample set problem. Considering that not all feature elements are related to the similarity estimation, the boosting process in the proposed method provides a good generalization of the feature elements that are most important in classification context. It also has the dimension reduction effect which may be very useful when the original feature dimension is high. The automatic metric adaptation and element selection in our boosted distance metric bridge the gap between the high-level similarity concept and low-level features. The experimental results have proven our proposed method is more effective and efficient than traditional distance metrics. In the future we would like to incorporate our new metric into state-of-the-art classification techniques and evaluate the performance improvement.

Acknowledgement: This work was supported in part by the Army Research Office (ARO) grant under W911NF-05-1-0404, and by the Center of Infrastructure Assurance and Security (CIAS), the University of Texas at San Antonio. The work of Nicu Sebe was done within the MUSCLE-NOE.

References

[1] M. Zakai, "General distance criteria," *IEEE Trans. on Information Theory*, pp. 94-95, January 1964.

[2] N. Sebe, M.S. Lew, D.P. Huijsmans, "Toward Improved Ranking Metrics", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, pp. 1132-1141, October, 2000

[3] M. Swain and D. Ballard, "Color indexing," *Intl. Journal Computer Vision,* vol. 7, no.1, pp. 11-32, 1991.

[4] R. M. Haralick, et al, "Texture features for image classification," *IEEE Trans. on Sys. Man and Cyb.*, 1990.

[5] B. M. Mehtre, et al., "Shape measures for content based image retrieval: a comparison", *Information Proc. Management*, 33(3):319-337, 1997.

[6] J. Amores, N. Sebe, P. Radeva, "Boosting the Distance Estimation: Application to the K-Nearest Neighbor Classifier", *Pattern Recognition Letters*, Vol. 27, No. 3, pp. 201-209, February 2006.

[7] R. Haralick and L. Shapiro, *Computer and Robot Vision II*, Addison-Wesley, 1993.

[8] R. E. Schapire, Y. Singer, "Improved boosting using confidence-rated predictions," *Machine Learning* 37 (3) (1999) 297–336.

[9] C. Domeniconi, J. Peng, D. Gunopulos, "Locally adaptive metric nearestneighbor classification," *IEEE Trans. PAMI* 24 (9) (2002) 1281–1285.

[10] J. Peng, et al., "LDA/SVM driven nearest neighbor classification," *IEEE Proc. CVPR*, 2001, pp. 940–942.

[11] E. P. Xing, A. Y. Ng, M. I. Jordan, S. Russell, "Distance metric learning, with application to clustering with side-information.", *Proc. NIPS*, 2003, pp. 505–512

[12] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, "Learning distance functions using equivalence relations," *Proc. ICML,* 2003, pp. 11–18.

[13] T. Hertz, A. Bar-Hillel, D. Weinshall, "Learning distance functions for image retrieval," *IEEE Proc. CVPR*, 2004, pp. 570–577.

[14] P. J. Huber, *Robust Statistics*, John Wiley & Sons, 1981.

[15] L. Tang, et al., "Performance evaluation of a facial feature tracking algorithm," *Proc. NSF/ARPA Workshop: Performance vs. Methodology in Computer Vision*, 1994.

[16] Y. LeCun, et al., MNIST database, http://yann.lecun.com/exdb/mnist/.

[17] A. Martinez, R. Benavente, *"The AR face database,"* Tech. Rep. 24, Computer Vision Center (1998).

[18] J. Matas, et al. , "Comparison of face verification results on the xm2vts database," *Proc. ICPR*, pp. 858–863, 1999.